Division of
CANCER PREVENTION

FRED HUTCHINSON
CANCER RESEARCH CENTER
A LIFE OF SCIENCE

EDRN Knowledge Environment

RDF Specification DRAFT

NASA Jet Propulsion Laboratory 4800 Oak Grove Drive Pasadena CA 91109-8001 http://www.jpl.nasa.gov/

Signature Page

This is an *official* document representing the policies, recommendations, technical considerations, and other aspects of the Early Detection Research Network's software products as codeveloped by EDRN's Data Management and Coordinating Center and NASA's Jet Propulsion Laboratory in its role as Informatics Center. As such, the following signatures signify *acceptance* and *endorsement* of this document.

Sudhir Srivastava, Ph.D., M.P.H., Chief, Cancer Biomarkers Research Group

Ziding Feng, Ph.D., EDRN Principal Investigator, DMCC

Daniel Crichton, M.S., EDRN Principal Investigator

Table of Contents

Introduction

The Early Detection Research Network (EDRN) researches and develops biomarkers and technologies for the clinical application of early cancer detection strategies. EDRN is an initiative of the National Cancer Institute (NCI) and brings together dozens of institutions to help accelerate the translation of biomarker information into clinical applications and to evaluate new ways of testing cancer in its earliest stages and for cancer risk.

Since its founding in April 2000, EDRN has fostered a highly collaborative, multiple disciplinary research venue to improve its techniques and methods. *Informatics* has been a core element of this collaboration. Computing technology and dense data dissemination have enabled the sharing, discovery, correlation, and synthesis of cancer research knowledge in efficient and often novel ways.

The Early Detection Research Network Third Report, March 2005.

The Informatics Center of EDRN is deploying a specialized *knowledge environment application* that enables EDRN members, other cancer researchers, doctors and clinicians, cancer sufferers and advocates, as well as the general public to explore, learn about, annotate, update, refine, consolidate, coalesce, and discover the collective data, information, and knowledge of EDRN. We call this application the EDRN Knowledge Environment.

A key component of the EDRN Knowledge Environment is an *import function*—that is, the ability to bring new and updated information into the system that hones and completes the total knowledge picture available. Repeated "injections" of such data from approved sources lead to a better and more refined picture of the communal wisdom of EDRN.

This document serves to specify the data format used by the EDRN by the EDRN Knowledge Environment.

Intended Audience

XXX.

Knowledge of RDF is essential towards understanding this specification, including the concepts of RDF (resources and properties) and the technologies used within RDF: XML, URI, URN, and URL. Knowledge of UML may be helpful as well.

Organization of this Document

This document contains the following sections:

XXX.

Object Model

In this section, we explore the classes and objects that comprise the information model for the EDRN Knowledge Environment. This section demonstrates the high-level relationships between the entities of the environment in order to comprehend the total information model of EDRN. The next section, XXX, details the properties of each of the entities in detail. Understanding of the Unified Modeling Language (UML) diagramming notation is helpful.

Unified Modeling Language (UML) Version 1.4.2, ISO/IEC 19501:2005.

Inheritance of Classes

Many of the entities within the EDRN Biomarker Database are identified; that is, they have a unique URI that is either assigned by software or defined manually. As such, most of the resources within the Biomarker Database have as an ultimate parent class a class called Identified Object, whose sole property is the resource's URI. The following UML class diagram depicts the inheritance relationships that trace ancestry to Identified Object:

Berners-Lee, *Universal Resource Identifiers, RFC-1630.* IETF 1994.

For example, as shown in the diagram, a Cancer is a kind of Disease, which is a kind of Identified Object. Likewise, an EDRN Biomarker Study is a kind of EDRN Protocol, which is also an Identified Object.

Relationships Between Objects

In addition to inheritance, there are composition and association relationships that further refine the biomarker object model. This section describes these relationships.

Registries and Registry Entries

Registries are merely containers of URIs. They track merely pairs of string identifiers to either URN or URL. As such, they are general purpose, and can be used as registries for biomarkers, specimens, publications, people—anything that has a URI.

There are three kinds of registries defined thus far:

- Swisspro Registry
- Unipro Registry
- EDRN Registry

A registry, as shown in the class diagram at right, contains zero or more Registry Entries.

People

People from all walks of life are modeled using a hierarchy of more and more specific classes as shown above in the inheritance class diagram. People also have relationships to other entities as shown at right.

As shown in the diagram, publications can have people as authors. Generic people, registered people, staff members, investigators, and principals investigators can all be authors.

EDRN protocols are discussed in publications (shown by the bent arrow).

Finally, only EDRN Biomarker Studies are associated with Principal Investigators. Note that in the real world principal investigators are of course associated with many other kinds of studies and protocols, however for purposes of the Biomarker Database, the model only shows biomarker studies as associated with them.

Note: investigators may take on several roles (principal investigator on one project, co-investigator on another, for example). The current class structure does not support such concepts, and may have to be changed.

Biomarkers

Biomarkers are of course the key area of focus for the EDRN Biomarker Database. Biomarkers therefore have a fairly complicated structure which we'll break down in two parts: their relationship to other classes and their relationship to themselves.

Associations with Cancer, Studies, and Resources

The following class diagram shows how Biomarker Research objects related to other classes within the model:

As shown, EDRN Biomarker Studies have the job of studying zero or more Biomarker Research objects. Such objects may be related to generic Resources, which are identified objects with a set of standard Dublin Core metadata.

Further, Biomarker Research objects have the job of detecting zero or more kinds of Cancer objects, which are kinds of Disease objects. Diseases affect organs.

Composite Biomarkers

Biomarker Research may also be developed into panels of biomarkers, that is, a single research effort into biomarkers that itself contains separate biomarkers. The diagram at right depicts this relationship.

Also shown are specific kinds of biomarker research. Currently, there is just one specialized kind: a molecular biomarker. This kind itself is further subclassed into protein molecule biomarker research and genetic molecule biomarker research.

Dublin Core Metadata Element Set. 1.1, Dublin Core Metadata Initiative 2006.

Resource Description Format

This section describes the data format used to import information about biomarkers from the EDRN Biomarker Database (or other compatible application) and into the EDRN Knowledge Environment. Note that this format is based largely in part on the current capabilities of the EDRN Biomarker Database, which currently is a subset of the object model in use by the EDRN Knowledge Environment. As such, the information in this section is seriously subject to change.

Namespaces and Schema

We interchange biomarker data by taking advantage of the Resource Description Format, or RDF. RDF is a technical recommendation of the World Wide Web Consortium and is an integral part of the Semantic Web. With RDF, we can describe all of the classes that comprise the Biomarker object model.

http://www.w3.org/RDF/

The RDF namespace URI for the Biomarker Database is urn:bmdb: (this will likely change to an URL in the future). The RDF Schema for the Biomarker Database is located at http://edrn.nci.nih.gov/xml/rdf/schema/biodb-0.0.0#.

The EDRN Knowledge Environment accepts imports of biomarker information expressed in schema-compliant RDF at the address http://edrn.nci.nih.gov/biodb-load-rdf. RDF may be presented in RDF/XML or N3 formats. Note that administrator permissions are currently required in order to load biomarker information.

Resource Classes and Properties

This section details each resource class defined in the schema and expected by the EDRN Knowledge Environment. For clarification of each class, see the Object Model section.

Biomarker Research

Biomarker Research objects use the RDF resource class URI of urn:bmdb:BiomarkerResearch.

Properties

Property URI	Value	Usage
urn:bmdb:MarkerType	Blank	Not currently used
urn:bmdb:MarkerId	UUID	Ignored
urn:bmdb:MarkerTitle	String	The title of the biomarker research
urn:bmdb:MarkerRegistry	String	Ignored
urn:bmdb:MarkerDescription	String	Summary or abstract of the biomarker research
urn:bmdb:RelatedPublication	urn:bmdb:Publication	Publications that describe the progress of the research
urn:bmdb:ClinicalDiagnosis	String	Ignored

Example

The following RDF/XML serialization describes a single biomarker for PSA using the SELDI technology:

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:bmdb="urn:bmdb:">
        <bmdb:BiomarkerResearch
                rdf:about="http://edrn.nci.nih.gov/bmdb/viewMarker.jsp?markerId=43940">
                <bmdb:MarkerType/>
                <bmdb:MarkerId>43940</bmdb:MarkerId>
                <bmdb:MarkerTitle>SELDI Protein Profiles</bmdb:MarkerTitle>
                <bmdb:MarkerDescription>
                        Protein profiles from two independent laboratories
                        using SELDI-TOF-MS and IMAC Proteinchip, using serum
                        from 500 prostate cancer cases (250 with
                        aggressive/advanced disease and 250 with intermediate-
                        or low-risk disease) as well as 250 prostate cancer
                        controls (biopsy-negative with a range of PSA values).
                        An additional group of 50 patients with other cancers
                        but no evidence of prostate cancer and 50 patients
                        with various inflammatory diseases will also be
                        examined.
                </bmdb:MarkerDescription>
                <bmdb:RelatedPublication
                        rdf:resource="http://edrn.nci.nih.gov/bmdb/viewMarker.jsp?markerId=43941"/>
        </bmdb:BiomarkerResearch>
</rdf:RDF>
```

Publication

Publication objects use the RDF resource class URI of urn:bmdb:Publication.

Properties

Property URI	Value	Usage
urn:bmdb:PublicationId	UUID	Ignored
urn:bmdb:PublicationTitle	String	Title of the article, paper, book, etc.
urn:bmdb:PubMedId	String	ID as assigned by PubMed
urn:bmdb:PublicationAbstract	String	Summary of the publication
urn:bmdb:PublicationContact	String	Summary or abstract of the biomarker research
urn:bmdb:PublicationContact	String	Affiliation of the primary author of the publication.

Example

The following RDF/XML serialization describes a single publication about α-methlacyl-Coa racemase:

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:bmdb="urn:bmdb:">
        <bmdb:Publication rdf:about="http://edrn.nci.nih.gov/bmdb/viewMarker.jsp?markerId=43941">
                <bmdb:PublicationId>43941</bmdb:PublicationId>
                <bmdb:PublicationTitle>
                        Humoral immune response to α-methylacyl-CoA racemase and prostate cancer
                </bmdb:PublicationTitle>
                <bmdb:PublicationPubMedId>15173267.0</bmdb:PublicationPubMedId>
                <bmdb:PublicationAbstract>
                        Although prostate-specific antigen (PSA) is a prototypic biomarker for prostate
                        cancer, it has poor specificity. Expression of α-methylacyl-CoA racemase
                        (AMACR), which is involved in the conversion of R-stereoisomers of
                        branched-chain fatty acids to S-stereoisomers, has been shown to be specifically
                        increased in prostate cancer epithelia. However, attempts to detect AMACR in
                        circulation have not been successful. Hence, we determined whether an immune
                        response to AMACR could be used as a serum biomarker for prostate
                        cancer.
                </bmdb:PublicationAbstract>
                <bmdb:PublicationContact>
                        Department of Pathology, University of Michigan Medical School,
                        Ann Arbor, MI 48109-0602, USA.
                </bmdb:PublicationContact>
        </bmdb:Publication>
</rdf:RDF>
```